Procesamiento del Lenguaje Natural, Revista nº 62, marzo de 2019, pp. 103-106 recibido 15-11-2018 revisado 07-01-2019 aceptado 30-01-2019

Similitud entre palabras: aportaciones de las técnicas basadas en bases de datos *

Word similarity: contributions of knowledge-based methods

Josu Goikoetxea Salutregi
Grupo Ixa, Universidad del País Vasco (UPV/EHU)
Manuel Lardizabal 1, 20018 Donostia
josu.goikoetxea@ehu.eus

Resumen: Tesis doctoral titulada "Hitzen arteko antzekotasuna: ezagutza-baseetan oinarritutako tekniken ekarpenak", defendida por Josu Goikoetxea Salutregi en la Universidad del País Vasco (UPV/EHU) y elaborada bajo la dirección de los doctores Eneko Agirre (Departamento de Lenguajes y Sistemas Informáticos) y Aitor Soroa (Departamento de Ciencias de la Computación e Intelingencia Artificial). La defensa tuvo lugar el 13 de julio del 2018 ante el tribunal formado por los doctores Arantza Diaz de Ilarraza (Presidenta, Universidad del País Vasco (UPV/EHU)), Diego Lopez de Ipiña (Secretario, Universidad de Deusto) e Itziar Aduriz (Vocal, Universidad de Barcelona). La tesis obtuvo la calificación de sobresaliente.
Palabras clave: Similitud semántica, redes neuronales, embeddings, multilingüe, bases de datos, corpus de texto

Abstract: Ph.D. thesis entitled "Hitzen arteko antzekotasuna: ezagutza-baseetan oinarritutako tekniken ekarpenak", written by Josu Goikoetxea Salutregi at the University of Basque Country (UPV/EHU) under the supervision of Dr. Eneko Agirre (Languages and Computer Systems Department) and Dr. Aitor Soroa (Computer Science and Artificial Intelligence). The viva voce was held on July 13 2018 and the members of the commission were Dr. Arantza Díaz de Ilarraza (President, University of Basque Country (UPV/EHU)), Dr. Diego Lopez de Ipiña (Secretary, University of Deusto) and Dr. Itziar Aduriz (Vocal, University of Barcelona). The thesis obtained excellent grade.
Keywords: Semantic similarity, neural networks, embeddings, cross-lingual, knowledge bases, text corpus

1 Introducción de la tesis

Esta tesis doctoral ha sido llevada a cabo en el grupo IXA de la Universidad del País Vasco (UPV/EHU) y su línea de investigación es la similitud semántica.

Desde la perspectiva de la psicología cognitiva, la similitud es una capacidad innata a los humanos, y sirve para estructurar la información proveniente de la realidad, para darla sentido. Tomando en cuenta esto último, es vital poder reproducir esa capacidad cognitiva en las máquinas si queremos que aprendan del mundo.

En concreto, la similitud se encuentra en el núcleo de muchas tareas del procesamiento del lenguaje natural, siendo crucial su correcta reproducción. En esta tesis se ha trabajado con modelos distribucionales que se basan en el teoría de Harris y que calculan representaciones densas (en inglés *embeddings*) de significados de palabras. Siguiendo el teoría de Harris, los *embeddings* de palabras que comparten contextos parecidos van a ser similares, y, por tanto, la capacidad de reproducir la similitud es intrínseca a esas representaciones. De hecho, una de las evaluaciones más conocidas es la similitud entre palabras: cuanto más se acercan los resultados de similitud de un modelo computacional a los criterios humanos, mejor es la calidad de sus representaciones.

La motivación principal de este trabajo es

*Esta tesis doctoral ha sido realizada con una beca predoctoral otorgada por el Vicerrectorado de euskera la Universidad Del País Vasco.

ISSN 1135-5948. DOI 10.26342/2019-62-13

mejorar la calidad de los *embeddings*, y así, mejorar también los resultados en similitud entre palabras. Teniendo en cuenta que los *embeddings* sólamente se basan en texto, y siguiendo la hipótesis de que la información semántica de métodos basados en corpus de texto y en bases de datos es complementaria, se han propuesto técnicas para combinar estas dos fuentes y crear representaciones híbridas de mayor calidad. Además, hemos extendido esos métodos al entorno bilingüe. Esta tesis consta de tres partes: 1) introducción de la tesis y precedentes de métodos basados en texto y bases de datos, 2) métodos propuestos para mejorar similitud semántica, y 3) conclusiones y trabajo futuro.

2 Estructura de la tesis

La primera parte se divide en los dos primeros capítulos. El primer capítulo está dedicado a definir la similitud semántica partiendo de una perspectiva cognitiva, para luego motivar el uso de la similitud en los modelos computacionales que calculan significados de palabras. Es importante mencionar la distinción que se hace entre similitud[1] y asociación[2], ya que en el procesamiento del lenguaje según la tarea nos interesa una u otra. El segundo capítulo resume los precedentes de la tesis: primero, presenta las bases teóricas de la Semántica Distribucional y los modelos distribucionales actuales basados en texto; luego, describe las características de los métodos basados en bases de datos; a continuación, explica los métodos que aúnan información de los dos anteriores; después, resume los principales métodos para calcular representaciones multilingües; finalmente, termina explicando el método de evaluación de las representaciones de palabras, que se realizará calculando la correlación Spearman entre los resultados de similitud del modelo computacional y los patrones-oro creados por humanos.

La segunda parte consta de tres capítulos, que coinciden con las tres publicaciones principales de la tesis. El tercer capítulo describe nuestra propuesta (Goikoetxea, Soroa, y Agirre, 2015) para crear *embeddings* de lexicalizaciones de bases de datos. Este método tiene dos fases: en la primera, se crea un seudocorpus con las lexicalizaciones emitidas en los caminos aleatorios llevados a cabo por la herramienta UKB[3]; en la segunda, se calculan los *embeddings* de las lexicalizaciones del seudocorpus mediante *word2vec*[4]. La base de datos que hemos usado es WordNet[5], evaluando los *embeddings* resultantes en un patrón-oro de similitud y otro de asociación, y comparando los resultados con métodos en el estado del arte.

En el cuarto capítulo presentamos ocho combinaciones (Goikoetxea, Agirre, y Soroa, 2016) que fusionan la información complementaria de espacios semánticos separados de texto y bases de datos que dan como resultado representaciones híbridas. Las combinaciones se dividen en cuatro grupos: combinaciones de *embeddings*, de corpus, de resultados y las basadas en correlaciones. Así como en el capítulo anterior, los resultados de las combinaciones se evalúan en siete patrones-oro (tres de similitud y cuatro de asociación), y se comparan con resultados de varios métodos en el estado del arte.

En el quinto capítulo extendemos todo lo explicado anteriormente a espacios bilingües (Goikoetxea, Soroa, y Agirre, 2018), y presentamos un algoritmo para crear seudocorpus bilingües a través de caminos aleatorios bilingües, un método para crear corpus híbridos bilingües y una adaptación de la arquitectura Skip-gram de *word2vec* que introduce restricciones bilingües a la función-objetivo original. Al igual que en los dos capítulos previos, evaluamos las representaciones en patrones-oro bilingües y comparamos los resultados con las representaciones en el estado del arte.

El capítulo final resume las principales contribuciones de las tesis y el trabajo futuro. En los apéndices se describe de forma detallada la arquitectura original de Skip-gram y la extensión de Skip-gram con restricciones propuesta en el capítulo cuatro.

3 Contribuciones más relevantes

En los siguientes apartados describimos las contribuciones en nuestra línea de investigación, así como los recursos generados.

[1] Incluye sinonimia, hiponimia e hiperonimia.

[2] Incluye, además de las de la similitud, meronimia, antonimia, asociaciones funcionales y otra serie de relaciones poco comunes.

[3] `http://ixa2.si.ehu.es/ukb/`

[4] `https://code.google.com/archive/p/word2vec/`

[5] En concreto, la versión 3.0 con glosas. Esta versión también se ha usado en los dos siguientes capítulos.

3.1 Seudocorpus de WordNet

El seudocorpus de Wordnet se ha creado a partir de UKB, una colección de herramientas basadas en grafos. Este tipo de técnicas tratan las bases de datos como si fueran grafos y procesan la totalidad de la estructura del grafo. Se ha modificado UKB de forma que en su algoritmo de caminos aleatorios emita a un fichero las lexicalizaciones de los conceptos por los que pasa dicho camino. De esta manera, se codifica la información estructural de WordNet en las coocurrencias de un seudocorpus, abriendo la posibilidad de procesarlo modelo distribucional.

3.2 Embeddings de WordNet

Se han procesado el seudocorpus de WordNet mediante Skip-gram y CBOW y calculado luego los *embeddings* de todas las lexicalizaciones. Si comparamos esas representaciones de WordNet con las de los métodos basados en bases de datos de ese momento, nuestras representaciones son mucho más compactas y computacionalmente menos costosas.

Los mejores resultados se han obtenido calculando los *embeddings* de WordNet con Skip-gram, superando los resultados del patrón-oro de similitud SimLex999 con un Spearman de 0.52 e igualando los del patrón-oro de asociación WordSim353 con un Spearman de 0.683 a los vectores de UKB del estado del arte. Además, se ha propuesto una técnica para combinar los resultados de diferentes fuentes de información, superando del estado del arte del momento con un Spearman de 0.552 en SimLex999 mediante la combinación de *embeddings* de WordNet, texto y vectores UKB.

Unido con la aportación del seudocorpus de WordNet, hemos comprobado que efectivamente las coocurrencias de caminos aleatorios codifican la información relacional de WordNet. Debido a esto, los *embeddings* de WordNet son capaces de recoger la misma información estructural que los vectores de UKB, pero en un formato mucho más compacto y eficiente.

3.3 Representaciones híbridas

Basándonos en la complementariedad de la información semántica en corpus de texto y bases de datos, se han propuesto métodos eficientes para crear representaciones híbridas combinando espacios semánticos separados de los dos recursos mencionados. De los ocho métodos propuestos, los mejores resultados son con la aplicación de PCA a la concatenación de *embeddings* de texto y de WordNet. Este última combinación introduce una ganancia absoluta media respecto a las correlaciones Spearman de *embeddings* de texto de 9.6 en los patrones-oro de similitud, 6.5 en los de asociación y 8.9 en general, sobrepasando con creces al método del estado del arte llamado *retrofitting*, que enriquece información *embeddings* de texto con información de WordNet. Además, se combinan hasta un total de seis recursos semánticos, superando con una combinación los seis recursos el estado del arte proveniente de diferentes métodos.

Por un lado, hemos demostrado que crear representaciones de significados a partir de espacios separados es más eficiente que crear representaciones que aprenden de forma conjunta desde esos espacios. Por otro, hemos comprobado que cuanto más recursos de diferente naturaleza semántica combinemos, mayor será la calidad de las representaciones.

3.4 Extensión bilingüe

En la ultima fase de la investigación se han extendido al entorno bilingüe las aportaciones descritas en los tres apartados anteriores. Hemos usado el inglés, el castellano, el euskera y el italiano, creando las siete combinaciones bilingües posibles en esos idiomas en tres tipos de corpus: texto, seudocorpus e híbrido. Además, hemos calculado todos esos *embeddings* usando tres métodos: el mapeo de espacios monolingües separados, procesamiento de corpus mediante Skip-gram original y con el Skip-gram que integra restricciones bilingües. Todos estos *embeddings* se han evaluado en tres patrones-oro bilingües creados automáticamente por nosotros. La combinación que mejores resultados ha obtenido es la de los *embeddings* calculados mediante el Skip-gram extendido procesando corpus híbridos bilingües, obteniendo resultados significativamente mejores que los *embeddings* obtenidos mediante el mapeo de espacios monolingües separados, e igualando a los vectores NASARI.

Las conclusiones obtenidas en los anteriores tres apartados se mantienen en espacios bilingües: a saber, que los seudocorpus bilingües de WordNet son capaces de codificar su información relacional, y que los *embeddings* de corpus híbridos bilingües son muy

eficientes en similitud. Además, hemos comprobado que la inserción de restricciones bilingües provenientes de WordNet mejoran todavía más los resultados en similitud.

3.5 Recursos

A lo largo de esta tesis se han hecho públicos para la comunidad científica los recursos descritos a continuación. En cuanto al software liberado, el algoritmo de caminos aleatorios para crear seudocorpus (tanto monolingües como bilingües) está implementado en la versión 2.1 de UKB[6], y la extensión de Skip-gram con restricciones se encuentra en el repositorio `github`[7]. Los *embeddings* monolingües[8] de WordNet y híbridos, así como todos bilingües[9] del apartado anterior también se han hecho públicos. En relación con esos últimos recursos bilingües, se puede acceder a todos los patrones-oro usados en el entorno bilingüe, incluyendo los dos patrones-oro monolingües de euskera creados expresamente para esa fase de la investigación.

Bibliografía

Goikoetxea, J., E. Agirre, y A. Soroa. 2016. Single or multiple? combining word representations independently learned from text and wordnet. En *AAAI*, páginas 2608–2614.

Goikoetxea, J., A. Soroa, y E. Agirre. 2015. Random walks and neural network language models on knowledge bases. En *Proceedings of the 2015 Conference of NAACL/HLT*, páginas 1434–1439.

Goikoetxea, J., A. Soroa, y E. Agirre. 2018. Bilingual embeddings with random walks over multilingual wordnets. *Knowledge-Based Systems*, 150:218–230.

[6] `http://ixa2.si.ehu.es/ukb/ukb_2.1.tgz`

[7] `https://github.com/JosuGoiko/word2vec_constraints`

[8] `http://ixa2.si.ehu.es/ukb/`

[9] `http://ixa2.si.ehu.es/ukb/bilingual_embeddings.html`